An Assurance Case Pattern for the Interpretability of Machine Learning in Safety-Critical Systems

Francis Rhys Ward✉ and Ibrahim Habli

Assuring Autonomy International Programme, The University of York, U.K.
{Rhys.Ward, Ibrahim.Habli}@york.ac.uk

Abstract. Machine Learning (ML) has the potential to become widespread in safety-critical applications. It is therefore important that we have sufficient confidence in the safe behaviour of the ML-based functionality. One key consideration is whether the ML being used is interpretable. In this paper, we present an argument pattern, i.e. reusable structure, that can be used for justifying the sufficient interpretability of ML within a wider assurance case. The pattern can be used to assess whether the right interpretability method and format are used in the right context (time, setting and audience). This argument structure provides a basis for developing and assessing focused requirements for the interpretability of ML in safety-critical domains.

Keywords: Interpretability · Explainability · Machine Learning · Artificial Intelligence · Assurance · Safety · Safety-Case

1 Introduction

Machine Learning (ML) algorithms are powerful tools and have applications in domains in which safety is a concern. One potential weakness of these algorithms is that they are often too complicated to understand - they may relate thousands of variables into patterns which cannot be understood by a human. This property is often referred to as the black-box problem. How can we accept these algorithms into safety-critical decision-making roles if we cannot understand how their decisions are made? [5, 13] This problem has limited the growth of ML algorithms in areas such as healthcare [20, 31].

A solution to this issue is to use ML algorithms which are more interpretable, or to try to explain their behaviour. In some sense an algorithm is interpretable if we can understand how it works and/or why it makes the decisions that it does make. [8] defines interpretability in the context of ML as 'the ability to explain or to present in understandable terms to a human' but notes that what constitutes an explanation is not well-defined. In practice, the term interpretability is used to refer to a number of distinct concepts [21]. A ML model may be said to be interpretable if the algorithm is simple enough for us to understand, otherwise there may be some post-hoc methods which can be used to interpret a black-box.

From a safety perspective, interpretability may help us to (1) understand the system retrospectively, i.e. to understand, with respect to a harm-causing action or decision, what went wrong, and why and (2) understand the system prospectively, i.e. to predict, mitigate, and prevent future harm-causing actions or decisions. But to what extent does machine learning need to be interpretable to provide assurance? To answer this question, we must decide on who needs to understand the system, what they need to understand, what types of interpretations are appropriate, and when do these interpretations need to be provided.

To this end, we present an argument pattern, i.e. reusable structure, that can be used for justifying the sufficient interpretability of ML within a wider assurance case. Structured argumentation is well-established in the safety-critical domain as a means for communicating, justifying and assessing confidence in properties of interest (e.g. risk reduction and acceptability). The pattern presents an explicit argument that can be used to assess whether the right interpretability method and format are used in the right context (time, setting and audience). We show how our pattern can be instantiated for assuring the interpretability of a system of neural networks intended for retinal disease diagnosis.

The following section provides a background to ML interpretability. In Section 3 we present an argument pattern for assuring that ML systems are interpretable. Then in Section 4 we motivate the need for interpretability in safety-critical ML systems.

2 A Brief Overview of Interpretability

There is a wealth of literature on interpretability of ML and AI [21], covering a wide range of philosophical and psychological perspectives [1, 12, 23, 26]; the legal implications of (un)interpretable ML [4, 11, 30]; technical methods for interpreting different types of ML models [3, 14, 15, 17, 19, 22, 27, 28]; and further discussions which try to bring some clarity to the field [7, 20, 21, 29].

Lipton in [21] seeks to clarify the myriad different notions of interpretability of ML models in the literature - what interpretability means and why it is important. It is noted that interpretability is not a monolithic concept and relates to distinct ideas. The distinction is often made between methods which are intrinsically *transparent* and post-hoc methods that attempt to *explain* a model. We identify the following types of interpretability. A model/system is:

- **Transparent** if we understand **how it genuinely works** (mechanistically, at some level, for some part of the process). A transparent model is one which is inherently simple enough for humans to understand. For example, for a learned model, could a human take the inputs and generate the same outputs as the model (in reasonable time)?
- **Explainable** if we can understand **why** it makes the decisions that it does make by using some post-hoc analysis and/or *approximation*, covering:
  - **Global explainability** techniques which approximate the model with a simpler more transparent one. This simple approximate model is an explanation.

- **Local explainability** techniques which map inputs to outputs and identify important inputs. These help us to answer the question 'what were the important factors in this decision?'

We can categorize some of the features of these different types of interpretability. Transparency provides faithful representations of the model, whereas explainable methods are often approximations, or incomplete explanations. Hence, there is a spectrum which captures the level of fidelity of different types of interpretability. Some methods interpret the whole model (global) whereas some interpret individual decisions (local). Transparency can be seen as an intrinsic property of a model (it is either easy to understand or not, or some degree in between), whereas explainability techniques are post-hoc methods which require some extra effort to implement.

It may be impossible for some systems/models to be fully and completely interpretable. For instance, a neural network may have some local explainability in that we can map certain inputs to outputs. But this does not provide a complete picture of how the model works globally and it is not transparent. We are interested in sufficient levels of interpretability needed to assure safety in different contexts.

3 An Argument Structure for the Interpretability of ML

Table 1: Phased Safety-Argument Development Alongside ML Life-cycle

Safety-Argument Phase	ML Life-cycle Stage	Interpretability Needs
Preliminary	Data Management	Global/Local: Identify Weaknesses in Data
Interim 1	Model Learning	Global: Aid Model Design
Interim 2	Model Verification	Global/Local: Identify Weaknesses in Model
Operational	Model Deployment	Local: Understand Decisions

Safety arguments, or "safety cases", are a well-established method used to assure system properties in the field of safety engineering. [16] advocates a phased safety argument approach wherein a number of safety case versions are issued alongside the developing technology, enabling an evolving safety argument. This phased safety argument will inform, and be informed by, the development process. This can be combined with the ML life-cycle from [2], which discusses the assurance of the complex, iterative process starting with the collection of data used to train an ML model, and ending with the deployment of that model. A safety argument should evolve with the ML life-cycle, as in Table 1. Because of the cyclical nature of the ML life-cycle, interpretability at a later stage may bring to light flaws which can then be accounted for on the next iteration.

In Figure 2, we define an argument pattern that explicitly addresses the interpretability assurance considerations, i.e. primary claims, argument strategies

and evidence. The argument is represented using the pattern language of the Goal Structuring Notation (GSN) [16]. GSN is a graphical argumentation notation which explicitly represents the individual elements of a safety argument (claims, evidence, and context) and the relationships that exist between these elements. When the elements of GSN are linked together in a network they are described as a "goal structure". We draw heavily from [6] which presents a pattern for arguing the assurance of machine learning, with a focus on clinical diagnosis. The first step is to ask why the project needs interpretability and set the desired requirements that the project should satisfy (e.g. being able to investigate accidents see Section 4.1). Figure 1 shows a key for GSN.

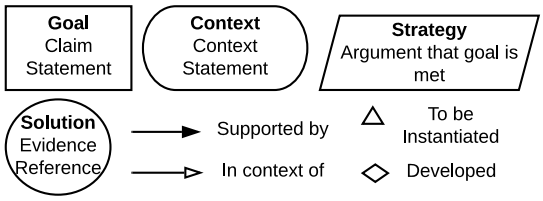

Fig. 1. GSN Key

- **Goal** - these are the claims being made in the argument.
- **Context** - the relevant additional information to the argument.
- **Strategy** - the argument approach for the support of a claim.
- **Solution** - evidence reference that claims have been met.
- **Supported by** - (solid arrows) indicates inferential/evidential relationships.
- **In context of** - (hollow arrows) declares contextual relationships.
- **To be instantiated** attached to an element indicates that some part of the element's content is a variable that requires instantiation. Variables are declared using curled braces, such as {ML Model}.

3.1 Interpretability Claim

In Figure 2, the starting point is the claim that the ML Model is sufficiently interpretable in the intended context. 'ML model', 'interpretable', and 'context' are variables in this claim to be instantiated As discussed in the previous section, the term 'interpretable' may refer to different types of interpretability. The substantiation of the 'ML model' will be the actual ML model being used, or a component of it, or the system as a whole - whatever needs to be interpreted. The context refers to the setting, time, and audience of the interpretation.

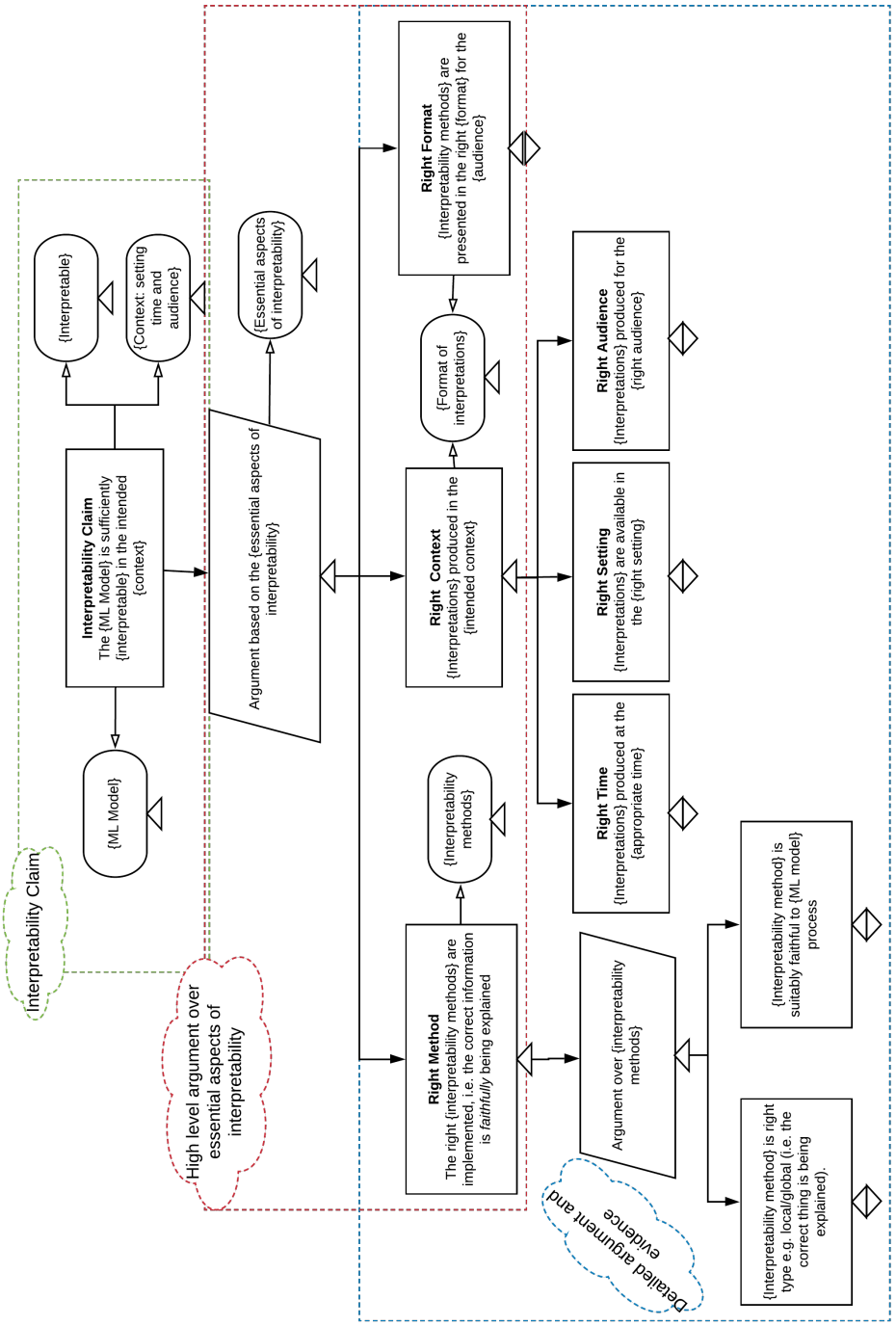

Fig. 2. General Argument Structure

3.2 High Level Argument

We identify three essential aspects of the interpretability argument, building on past work on context-aware systems [10]:

- Right Method - The right interpretability methods are implemented, i.e. the correct information is faithfully being explained.
- Right Context:
  - Time - Interpretations produced at the appropriate times.
  - Setting - Interpretations are available in the right setting.
  - Audience - Interpretations produced for the right audience.
- Right Format - The interpretability methods are presented in the right format for the audience.

A detailed argument over these essential aspects is presented in the next subsections.

3.3 Argument Over Interpretability Methods

This is the argument that suitable interpretability methods have been implemented, a method may simply be choosing a transparent model, or employing some post-hoc explainability techniques. There are two parts to this argument, first that the methods provide the type of interpretability required to satisfy the high level interpretability claim (e.g. if the claim is that the ML model is locally explainable in the context of accidents then the methods must provide this local explainability). Secondly the interpretability methods must be suitably faithful to the model process; these methods may be approximations to the model and may therefore not be accurate interpretations in all cases [29]. The interpretability methods must satisfy some desired level of fidelity in the given context. Both of these being satisfied equates to the correct information being explained.

Once a set of interpretability methods has been proposed, evidence that these methods are sufficient for purpose must be gathered. There are at least three different things which must be evaluated with regard to interpretability: how satisfying and appropriate produced interpretations are to stakeholders; how faithful interpretations are to the actual model workings; and the relevance of the interpretation being given. There is some initial research on how to evaluate the interpretability of ML models. [25] outlines how levels of explainability can be measured with respect to different user groups. [8] proposes an evidence-based taxonomy of evaluation approaches for interpretability. These are ways in which interpretations can be evaluated with respect to how effective they are at *convincing users*. Whilst it is important that stakeholders are satisfied with interpretations, these interpretations also need to be an accurate depiction of how the system acctually works.

Especially in safety-critical systems, it is important that interpretations, or explanations, of how a system works are not only convincing and satisfying but also reliably a faithful account of how the model is actually working. [28]

presents a technical method for evaluating the faithfulness of a certain kind of local explanation technique. These types of evaluation help users to understand how a model is genuinely working, even so far as the explanations can help users to gain enough insight to improve the model. [19] evaluates *fidelity* (faithfulness to the model) of explanations vs interpretability (how easy it is to understand) finding there are trade-offs between the two.

Recent work has highlighted the capacity of even high-fidelity explanations to mislead users [18]. Three key issues with current post-hoc methods, when optimised for fidelity, are described: i) they do not capture causal relationships; ii) they cannot choose between multiple (qualitatively different) high-fidelity explanations; iii) they can vary significantly with small perturbations of the input data. These problems lead to the possibility that current explainability techniques can actually mislead users. Importantly, explanations must also provide the most relevant information.

3.4 Argument Over Context

For simplicity we split context into time, setting, and audience.

- **Right Time:** Interpretations must be provided at the right time to avoid being intrusive or confusing. Not every decision may need to be explained and some interpretations may be needed in real time whereas others may only need to be produced under specific circumstances. For example, a diagnostic system may need to provide local explanations to clinicians alongside every diagnosis prediction, whereas an autonomous vehicle may only need to provide an explanation when an incident has occurred.
- **Right Setting:** It is important that interpretations are usefully available to the audience in the correct setting. Consider again a diagnostic tool, interpretations must be available to doctors in the clinical setting alongside diagnosis predictions. It is not useful for engineers to be able to produce interpretations if the audience do not have access to them in the relevant setting.
- **Right Audience:** Interpretations must clearly be provided to the right people to satisfy the interpretability claim and to satisfy the motivations for interpretability, e.g. policy makers vs developers vs users .

3.5 Argument Over the Format of Interpretations

The format of the interpretations is key. Once suitable methods for interpreting the system have been chosen, they must be presented in a format which is comprehensible and relevant to the audience. Section 3.3 discusses how to evaluate the extent to which interpretations are appropriate and satisfying to stakeholders and Section 4.4 outlines the needs of different stakeholders.

3.6 Example: Deep Learning for Diagnosis in Retinal Disease

We now examine a paper by DeepMind [9] that presents a system of two Neural Networks (NNs) working to predict retinal disease from scans of the eye. The paper purports to address the "black-box problem" by producing a midpoint result in the system. The first model takes as input a scan of the retina and produces a tissue-segmentation map. The second neural network takes the segmentation map and outputs a diagnosis and referral (with confidence levels). This process supplies some system-level transparency. We can instantiate this example in our argument structure as follows (Figure 3):

Interpretability Claim: The desired type of interpretability is transparency at the level of the system logic, the system being the combination of the two NNs. The context is defined by: the setting - the retinal diagnosis pathway; the time that interpretations are produced - alongside the system diagnosis prediction; the audience - the retinal clinicians.

Argument Over Method: The method by which interpretability is produced is that the system structure, including the production of the segmentation map, closely resembles the normal decision-process used by clinicians. This means that the system logic is inherently comprehensible, i.e. transparent, to the retinal clinicians. Note that this is true even though the individual NNs being used are not interpreted in any way. This is clearly a faithful method of interpreting the system logic, as transparency of the system is by definition faithful (the interpretation of the system logic is the system logic itself).

Argument Over Context: The audience are the retinal clinicians, and they need interpretations of system behaviour in the clinical setting and alongside each system diagnosis prediction.

Argument Over Format: The format of the interpretation is the transparent system logic, including the segmentation map. Presumably, the same prediction accuracy results could have been achieved without including the mid-point output of the segmentation map. Including this step allows clinicians to understand the system logic, since the production and use of the segmentation map are part of the normal clinical process and are understood by the retinal clinicians.

In summary, the healthcare setting here is clearly safety-critical and the designers of this system have identified interpretability as a requirement of the system in order that clinicians are able to understand and verify the system's predictions. Even though the individual NNs used were not interpreted, the method still provided some transparency of the system logic to the retinal clinicians, increasing their understanding of, and trust in, the ML system.

4 Discussion: Key Safety Interpretability Questions

4.1 Why Do We Need Interpretability in Safety-Critical Domains?

There are many reasons why we should want our ML systems to be interpretable. Interpretability may:

– Increase **insight** into model behaviour (and into the operational domain).

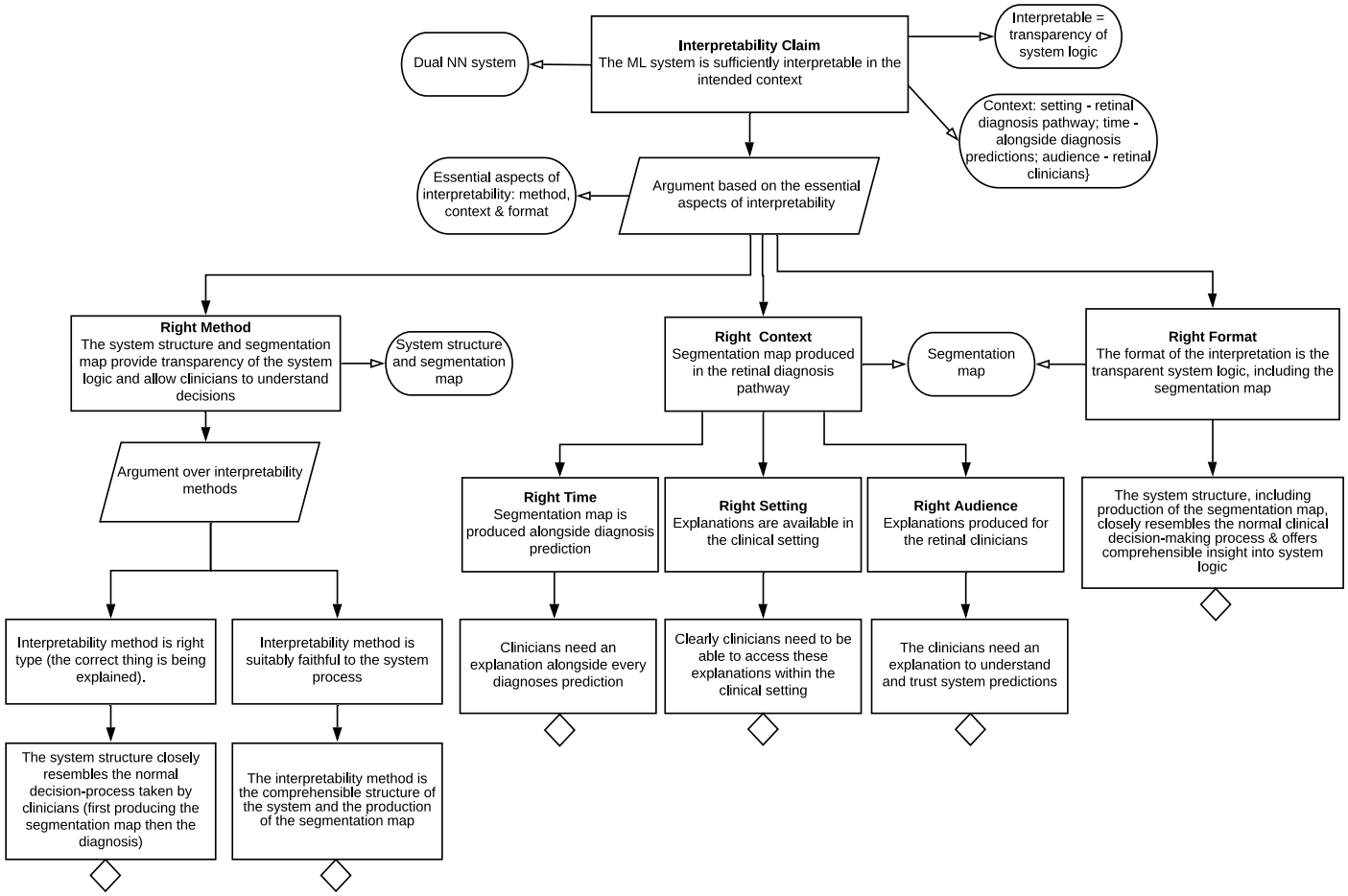

Fig. 3. DeepMind Example

- Identify **weaknesses** of the model, cases where the model under-performs.
- Enable the increase of **robustness** - i.e. assurance that the system will behave as intended in new environments/situations.
- Inform effective **improvements**/corrections.
- Protect against **unfair** models helping to avoid discrimination.
- Improve **trust** in the model and allow informed consent [31]

These advantages are beneficial in any domain of ML use. With regards to safety, interpretability is of interest for two key reasons:

- To understand the system retrospectively: to understand, with respect to a harm-causing action or decision, what went wrong, and why. This is important for post-hoc system diagnostics, establishing **accountability**, and accident inquiries.
- To understand the system prospectively: to **predict**, mitigate, and **prevent** future harm-causing actions or decisions.

Furthermore, the right to an intelligible explanation is supposedly required by law under the well-known 2018 GDPR regulation [11]. However, [24] argues that a right to explanation of automated decision-making does not exist in the GDPR due to the fact that the GDPR lacks precise language as well as explicit and well-defined rights and safeguards against automated decision-making. This closely relates to the lack of a precise language in the technical field of ML interpretbility [21].

4.2 What Needs to be Interpreted?

The different types of interpretability identified in Section 2 result in the interpretation of a set of distinct objects or processes. Transparency may refer to: the transparency of the whole model, wherein the entire global logic of the model can be explained and understood by a human; the transparency of the learning algorithm, we may understand that some algorithms converge to a solution in reasonable time (e.g. linear models), whereas we may not know whether another algorithm finds an optima at all (e.g. neural networks) [21]; transparency of parameters and model structures, do we understand what these are referring to and do they even map to human-understandable concepts? Similarly post-hoc explainablity methods may try to explain and interpret these processes, e.g. through approximating the global logic of a model, or they may explain local decisions. Global interpretability methods generate evidence that applies to a whole model (or system), and can be used to support safety assurance by allowing reasoning about all possible future outcomes. Local methods generate explanations for an individual system decision, and may be used to 1) predict how the system will behave in specific situations and 2) analyse why a particular problem occurred, and to improve the model so future events of this type are avoided.

4.3 When are Interpretations Needed?

Interpretations will be needed for different reasons during development and operation (Figure 4). ML developers may seek global explanations to better understand the model to aid design; stakeholders will need different types of interpretations during operation (local explanations may be more important during operation to explain individual cases - e.g. when explaining why an accident occurred). During development interpretations will be needed for:

- **Data Management** - interpreting the model may identify imbalances/gaps in the data.
- **Model Selection** - the interpretability of a model should influence this.
- **Model Learning** - being able to interpret the model will inform the model learning stage, e.g. in aiding hyper-parameter selection, data augmentation, etc.
- **Model Verification** - being able to interpret model decisions will aid verification and help to identify the cause of model weaknesses.

And during operation:

- **Normal operation** - e.g. for advisory systems such as diagnostic tools explanations may be compulsory.
- **In cases where the model is known to underperform** - which will aid contestability or identifying when to hand over control to a human.
- **Accident or incident Investigation** - Local explainability (e.g. counterfactual) to discover why particular decisions were made.
- **Model Run-time Improvement/Learning** - To improve models as new data and situations are encountered.

4.4 Who Needs an Interpretation?

Different stakeholders need different types of interpretations, consider lay users, expert users, designers, etc. Developers need explanations and transparency to understand how the model works in order to predict when undesirable model behaviour will occur and make corrections and improvements. Whilst developers may need some local explainability to understand and account for edge cases, in general they will need global interpretability to aid design. End-users will need local explanations to satisfy understanding of individual decisions. Figure 4 lists some potential stakeholders and the explanation needs for each.

5 Summary

In this paper, we built on previous work, which developed an assurance argument pattern for reasoning about ML in safety-critical domains. We extended this argument pattern by identifying interpretability as a key consideration. The extended argument pattern can be used to guide developers of ML systems as

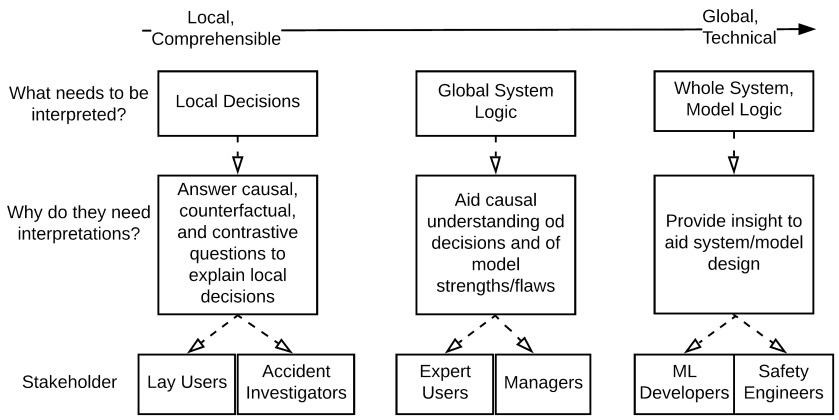

Fig. 4. Interpretation needs for different stakeholders

part of a wider safety or assurance case. It identifies how to create a structured assurance argument for the interpretability of ML models to support a decision over the deployment of the models in safety-critical applications. The key points in the argument are the essential aspects: right method, right context, and right format. These are claims that we have identified as necessary to form an explicit argument over interpretability; importantly these claims must also be supported by appropriate evidence.

The focus of future work should be to evaluate the applicability of the argument structure which should be presented to ML practitioners, their feedback should be used to make any necessary improvements. Further work may expand our argument structure to address different cases, for instance by drawing a more concrete link between relevant assurance properties and clear interpretability needs in a particular system. We hope that this argument structure will provide a clear basis for developing and assessing requirements for the interpretability of ML in safety-critical domains.

References

1. Achinstein: The nature of explanation (1983)
2. Ashmore, R., Calinescu, R., Paterson, C.: Assuring the machine learning lifecycle: Desiderata, methods, and challenges (2019)
3. Avati, A., Jung, K., Harman, S., Downing, L., Ng, A., Shah, N.H.: Improving palliative care with deep learning (2017)
4. Budish, R., et al.: Accountability of ai under the law: The role of explanation (2017)
5. Burton, S., Habli, I., Lawton, T., McDermid, J., Morgan, P., Porter, Z.: Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective (2020)

6. Chiara: A pattern for arguing the assurance of machine learning in medical diagnosis systems
7. Doran, D., Schulz, S., Besold, T.R.: What does explainable ai really mean? a new conceptualization of perspectives (2017)
8. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning (2017)
9. Fauw, J.D., et al.: Clinically applicable deep learning for diagnosis and referral in retinal disease (2018)
10. Fischer, G.: Context-aware systems: The 'right' information, at the 'right' time, in the 'right' place, in the 'right' way, to the 'right' person (2012)
11. Goodman, B., Flaxman, S.: European union regulations on algorithmic decision-making and a "right to explanation" (2016)
12. Grimm: The goal of explanation (2010)
13. Habli, I., Lawton, T., Porter, Z.: Artificial intelligence in health care: accountability and safety. Bulletin of the World Health Organization **98**(4), 251 (2020)
14. Hendricks, L.A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., and, T.D.: Generating visual explanations (2016)
15. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: Learning basic visual concepts with a constrained variational framework (2017)
16. Kelly, T.: A systematic approach to safety case management. (2003)
17. Koh, P.W., Liang, P.: Understanding black-box predictions via influence functions (2017)
18. Lakkaraju, H., Bastani, O.: How do i fool you?: Manipulating user trust via misleading black box explanations (2019)
19. Lakkaraju, H., Kamar, E., Caruana, R., Leskovec, J.: Interpretable explorable approximations of black box models (2017)
20. Lipton, Z.: The doctor just won't accept that! (2015)
21. Lipton, Z.C.: The mythos of model interpretability (2017)
22. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions (2017)
23. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences (2018)
24. Mittelstadt, B., Russell, C., Wachter, S.: Explaining explanations in ai (2018)
25. Mohseni, S., Zarei, N., Ragan, E.D.: A survey of evaluation methods and measures for interpretable machine learning (2018)
26. Mueller, S.T.: Explanation in human-ai systems: A literature meta-review synopsis of key ideas and publications and bibliography for explainable ai (2019)
27. Olah, C., Schubert, L., Mordvintsev, A.: Feature visualization how neural networks build up their understanding of images (2017)
28. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?" explaining the predictions of any classifier (2016)
29. Rudin, C.: Please stop explaining black box models for high-stakes decisions (2018)
30. Wachter, S., Mittelstadt, B., Floridi, L.: Why a right to explanation of automated decision-making does not exist in the general data protection regulation (2017)
31. Watson, D., et al.: Clinical applications of machine learning algorithms: beyond the black box (2019)